MathEdu Journal of the Korean Society of Mathematical Education Series A

Research Article

The Mathematical Education 2024;63(2):187-207
https://doi.org/10.7468/mathedu.2024.63.2.187
pISSN 1225-1380, eISSN 2287-9633

Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations

Byoungchul Shin[1], Junsu Lee[2]*, Yunjoo Yoo[3]

[1]Teacher, Suwon Foreign Language High School
[2]Teacher, Hwahong High School
[3]Professor, Seoul National University

ABSTRACT

In this study, we explored the feasibility of automatically scoring descriptive assessment items using GPT-4 based ChatGPT by comparing and analyzing the scoring results between teachers and GPT-4 based ChatGPT. For this purpose, three descriptive items from the permutation and combination unit for first-year high school students were selected from the KICE (Korea Institute for Curriculum and Evaluation) website. Items 1 and 2 had only one problem-solving strategy, while Item 3 had more than two strategies. Two teachers, each with over eight years of educational experience, graded answers from 204 students and compared these with the results from GPT-4 based ChatGPT. Various techniques such as Few-Shot-CoT, SC, structured, and Iteratively prompts were utilized to construct prompts for scoring, which were then inputted into GPT-4 based ChatGPT for scoring. The scoring results for Items 1 and 2 showed a strong correlation between the teachers' and GPT-4's scoring. For Item 3, which involved multiple problem-solving strategies, the student answers were first classified according to their strategies using prompts inputted into GPT-4 based ChatGPT. Following this classification, scoring prompts tailored to each type were applied and inputted into GPT-4 based ChatGPT for scoring, and these results also showed a strong correlation with the teachers' scoring. Through this, the potential for GPT-4 models utilizing prompt engineering to assist in teachers' scoring was confirmed, and the limitations of this study and directions for future research were presented.

Keywords Prompt engineering, Automatic wcoring, Descriptive qssessment, GPT-4 based ChatGPT, Generative artificial intelligence, Permutations and combinations

Received February 29, 2024; **Revised** April 1, 2024; **Accepted** May 3, 2024

***Corresponding author** Junsu Lee
E-mail jnsulee@gmail.com
2000 Mathematics Subject Classification 97C40

서 론

미래 교육에서는 교과서 지식을 암기하고 이를 선다형으로 평가하는 방식보다 다양한 자료를 비교, 분석해 문제를 해결하고, 이를 자신의 언어로 재구성해 설명할 수 있는 능력을 평가할 수 있는 서술형 평가가 더 중요해질 것으로 예상된다(Han & Koh, 2014; Kim & Lee, 2013; Park et al., 2023). 서술형 평가는 학습자의 수학 개념의 이해 및 기능의 숙달 여부 등에 대한 종합적인 정보를 확인할 수 있고, 결과를 학습자에게 환류함으로써 개인별 맞춤형 피드백을 가능하게 한다(Lee et al., 2021; Na et al., 2018). 이는 National Council of Teachers of Mathematics (2000)에서 평가는 수학 학습을 지원하고 교사와 학생 모두에게 유용한 정보를 풍부하게 제공해야 한다고 강조한 바와 일맥상통한다. 하지만 Noh 외 (2008), Chung 외 (2012), Kim 외 (2014)는 서술형 평가의 시행이 교사에게 시간적, 인지적 부담을 부과하고 있음을 지적하였으며, Pang 외 (2023)는 수업 중 짧은 시간 안에 학생들의 풀이 과정을 모두 점검하고 피드백 하는 것은 어려움이 있다고 하였다. 더불어 채점의 신뢰도 확보를 위해 거치는 채점 기준 마련, 가채점, 채점 기준 적용, 채점 기준 조정, 재채점의 반복 과정이 교사에게 상당한 부담을 줄 것이라 예상된다. 하지만 Sung (2023)의 연구에서 인공지능 교육시스템을 통해 학생들의 답안을 즉각적으로 채점하고 피드백을 제공한 결과, 이는 학습자에게 긍정적인 영향을 끼친다고 볼 수 있었기 때문에 채점 과정을 자동화하여 교사의 평가를 보조하는 것에 대한 필요성을 논할 수 있다.

이러한 문제의식으로부터 서술형 평가에서 채점 단계를 자동화하고자 하는 시도는 과거부터 존재해 왔다. 우리나라에서는 한국교육과정평가원 주도로 다양한 서술형 평가 자동 채점 연구가 진행되었다. 서답형 문항 자동 채점 프로그램 도입 방안 연구(Jin et al., 2006, 2007, 2008), 대규모 평가를 위한 서답형 문항 자동 채점 프로그램 개발(Noh et al., 2012, 2013, 2014), 한국어 문장 수준 서답형 자동 채점 프로그램 개발(Noh et al., 2015, 2016), 컴퓨터 기반 서·논술형 자동 채점 방안 설계(Park et al., 2022, 2023) 등이 그 예이다. 하지만 이러한 연구들에서 논의한 방법론은 몇 가지 한계를 가진다. 먼저 채점 정확도를 높이기 위해 지도 학습(Supervised learning)[1] 을 채택하고 있지만 이는 상당한 양의 기채점 데이터를 필요로 한다. 또한, 선행 연구들에서 형태소 기반의 자연어 처리(Natural Language Processing, NLP)[2] 기술을 통한 한국어 전처리를 활용하고 있는데 이는 본 연구에서 논의하고자 하는 수식을 포함하고 있는 수학 교과에서 잘 작동하지 않을 수 있다.

한편 2022년 11월 OpenAI에서 발표한 GPT-4 (Generative Pre-trained Transformer-4) 기반의 ChatGPT를 기점으로 전 세계가 본격적인 인공지능(Artificial Intelligence, AI) 시대에 들어오면서(Jung et al., 2023) 생성형(Generative) AI를 활용한 자동 채점에 대한 새로운 접근 방식이 주목받고 있다. GPT-4는 방대한 양의 데이터를 사전 학습하여 현존하는 거대 언어 모델(Large Language Model, LLM) 중 SOTA (State-Of-The-Art) 모델로 평가받는다. 특히 비지도학습의 Zero-Shot 러닝이 가능하고, 준수한 수학적 이해 역량을 가진다는 점에서 기존의 한계를 극복한 자동 채점에서의 활용가능성을 나타내고 있다. 뿐만 아니라 ChatGPT와 같은 챗봇(Chatbot) 서비스를 제공하여 사용자가 편리하게 입력값에 대한 출력값을 얻을 수 있어 학교 현장에서의 활용 가능성 또한 보여주고 있다.

ChatGPT를 활용한 서술형 자동 채점 방안에 대한 초기 연구가 지리 및 과학 교과에서 이루어졌다. Baek 외 (2023)는 과학 교과 서술형 평가에서 GPT-3.5 기반의 ChatGPT의 채점 결과와 교사와의 채점 결과를 비교한 결과 낮은 상관관계가 나타났고, Seong과 Shin (2023)은 지리 교과 서술형 평가에서 일부 서술형 문항에 대해 GPT-4 기반의 ChatGPT와 교사가 부여한 점수 사이에 유의미한 상관관계를 얻었다. 두 연구 모두 ChatGPT에 입력하는 명령어를 구조화하여 채점을 수행했지만 채점 결과가 일관적이지 않거나 피드백이 불완전하는 등의 한계를 지적하였고, 보다 정교하고 정확한 결과를 얻기 위한 후속 연구로 프롬프트 엔지니어링(Prompt Engineering)를 활용한 채점 시도가 필요하다고 하였다.

프롬프트(Prompt)란 사용자가 AI에게 입력하는 명령어(Giray, 2023)를 뜻하고, 프롬프트 엔지니어링이란 AI 모델의 입력 프롬프트를 설계, 개선 및 최적화하는 프로세스를 의미한다(Ekin, 2023). ChatGPT는 내부 구조가 공개되지 않은 엔드 투 엔드(End-to-End) 모델이므로 사용자가 내부 구조를 원하는 방향으로 수정하기 어렵기 때문에 입력 변수를 조정하는 방법을 고려할 수 있으며, 프롬프트 엔지니어링이 그 방법 중 하나이다. Wei 외 (2022)는 생각의 사슬(Chain of Thought, CoT) 프롬프트 기법을 제시하였으며, 중간 단계를 생성하도록 지시하면 복잡한 추론의 작업을 향상할 수 있다고 하였다. Kojima 외 (2022)는 CoT 프롬프트 기법을 개선한 자기 일관성(Self-Consistency, SC), 생각의 나무(Tree Of Thought, ToT) 프롬프트 기법을 사용하면 수학적 사고 및 추론 능력이 높아질 수 있다고 하였다. 본 연구에서는 다양한 프롬프트 엔지니어링 기법을 적용하여 ChatGPT의 복잡한 추론 등의 능력을 향상시키고, 이를 서술형 평가 자동 채점에 적용함으로써 그 결과에 대해 논의해보고자 한다.

구체적인 연구 질문은 다음과 같다.

1. 프롬프트 엔지니어링을 통한 GPT-4 기반의 ChatGPT의 순열과 조합 서술형 문항 자동 채점 결과는 어떠한가?
2. 프롬프트 엔지니어링을 통한 GPT-4 기반의 ChatGPT의 순열과 조합 서술형 문항 자동 채점 결과와 교사의 채점 결과의 차이는 어떠한가?

이론적 배경

1. 수학과 평가

(1) 서술형 평가의 특징

2022 개정 수학과 교육과정에 따르면 수학과 평가의 방향은 학생의 수학 학습에 대한 정보를 수집 및 활용하여 학생의 주도적 학습과 성장을 지원하는 것으로 평가의 과정을 중시하며 수학 내용 체계의 지식 • 이해, 과정 • 기능, 가치 • 태도의 균형 있는 평가를 강조한다. 특히, 문제해결, 추론, 의사소통 역량과 같은 수학 교과 역량을 평가함에 있어서 수학의 개념, 원리, 법칙을 문제 상황에 맞게 활용하여 적절한 해결 전략을 탐색하여 문제를 해결하고 반성하는지, 논리적인 절차를 수행하고 추측의 근거를 제시하는지, 수학 용어, 기호, 표, 그래프 등의 수학적 표현을 정확하게 사용하는지 등을 고려한다(Ministry of Education, 2022). 이러한 역량을 평가하기 위해 2009 개정 수학과 교육과정부터 선택형 위주의 평가보다 서술형 평가가 강조되고 있다(Han & Koh, 2014; Kim & Lee, 2013). 서술형 평가는 주어진 문제를 해결하는데 필요한 다양한 종류의 지식을 바탕으로 해결 방안을 구안, 검증, 분석 등의 고등 정신 기능들을 평가하는데 유리하며 문제해결 과정을 올바르게 이해하고 있는지 파악하는 평가유형이다(Chang & Kim, 2014; Kim et al., 2012). 서술형 평가를 통해 학생들이 어떻게 문제를 해결하고 추론하는지 분석함으로써 수학적 개념의 이해 수준과 수학적 기능의 숙달 정도, 문제 해결 전략과 방법, 오개념 등을 파악하여 학습자에 대한 구체적인 정보를 얻을 수 있다(Na et al., 2018). 또한 교사는 서술형 평가를 통해 학생들의 문장 형태의 답안에 대해 채점 및 결과를 환류함으로써 교사가 설계한 교수학습과 연계하여 학습 내용에 대한 이해도를 파악하고 이를 토대로 개인별 맞춤형 피드백을 할 수 있다(Lee et al., 2021).

서술형 평가시 문항을 채점하는 방법에는 총체적 점수화 방법과 분석적 점수화 방법이 있다. 총체적 점수화 방법은 문제의 해결 과정 전체를 종합하여 단일 점수로 평가하는 방법이고, 분석적 점수화 방법은 문제 해결 과정 및 단계를 구체화하여 각 단계별로 채점 요소를 세우고 점수를 부여하는 방법이다. 분석적 점수화 방법은 학생의 단계별로 수치화 된 점수를 부여함으로써 총체적 점수화 방법보다 채점자 간의 평점 차를 줄이고 동일한 채점자 내에서도 일관성 및 객관성을 유지할 수 있는 장점이 있다(Hwang et al., 2012). Kim과 Lee (2013)에 따르면 189명의 중학교 수학 교사들을 대상으로 한 설문에서 분석적 점수화 방법은 수행 판단 준거에 대한 정보를 제공하며 학생들에게 결과에 대한 공정성을 제공할 수 있어 85.2%의 교사가 서술형 문항 채점시 분석적 점수화 방법을 선호하였다. 하지만 서술형 문항의 채점에는 같은 답안에 대한 교사들의 채점 결과가 다를 수 있다는 단점이 있다(Na et al., 2018; Seo et al., 2010; Seong & Kwon, 1999).

이러한 이유로 Lee 외 (2021)는 서술형 문항 제작 원리 중 하나로 '채점 기준을 마련할 때, 가채점을 통하여 미리 만든 채점 기준을 적용해 본 뒤 채점 기준을 조정'할 것을 권고하였다. 출제자의 예상과 달리 서술형 문항은 학생들의 다양한 답안이 나올 수 있기 때문에 채점기준으로 채점하기 어려운 상황이 발생할 수 있기 때문이다. 따라서 교사는 서술형 평가의 채점을 위해 채점 기준 마련, 가채점, 채점 기준 적용, 채점 기준 조정, 재채점의 반복 과정을 거쳐야 한다는 것이다. 실제로 Noh 외 (2008)의 연구에 참여한 수학 교사 116명 중 83명이 한 문항당 4회 이상 채점을 실시하고 있다고 응답하였다. Noh 외 (2008), Chung 외 (2012), Kim 외 (2014)등의 연구에 참여한 교사들 중 상당수는 '채점의 부담', '문항 개발의 어려움' 등 채점과 관련된 문제 해결에 대한 요구가 있었다.

이러한 문제의식으로부터 서술형 평가에 대한 채점의 일부 또는 전부를 자동화하고자 하는 시도가 다양하게 진행되어 왔다. 한국교육과정평가원에 의해 자동 채점 연구가 시도되어 왔고(Noh et al., 2012, 2013, 2014, 2015, 2016; Park et al., 2022, 2023), 특히 Park 외 (2023)의 연구에서는 한국교육과정평가원 최초로 수학 서술형 평가에 대한 자동 채점을 시도하기도 하였다.

(2) 순열과 조합 단원의 평가

학교 수학의 한 분야인 조합론은 모든 가능한 경우를 고려하면서 각 상황의 경우를 계산하는 과정에서 명확한 사고능력

과 조합론적 태도를 기르는데 도움을 주며, 학생들은 논리적 정당화, 추상화 등의 과정을 경험할 수 있다(Kapur, 1970; Sriraman & English, 2004). 하지만 전통적으로 고등학교 순열과 조합 단원은 교사가 가르치기 어려운 단원 중 하나이고(Kim et al., 2009a), 학습자는 순열과 조합 단원에서 '순서에 관한 오류', '중복에 관한 오류', '대상의 구별에 관한 오류' 등 세기 과정에서 다양한 오류를 보이며 '합의 법칙', '곱의 법칙'에서 발생되는 인식론적 장애로 인해 세기 문제가 어렵다고 생각한다(Batanero et al., 1997; Choi & Cho, 2016; Kim et al., 2007). 순열과 조합 단원에서 이러한 어려움으로 인해 '구조적 동형을 활용한 교수 방안', '발견을 통한 지도방안' 등 교수 학습 방안에 대한 연구가 활발하게 진행되기도 하였다(Kim et al., 2009b; Kim et al., 2011). 특히, 순열과 조합 단원에서 서술형 평가를 통해 학생들의 풀이 과정, 표상의 사용 등을 종합적으로 분석할 필요가 있고(Choi & Cho, 2016), 이를 피드백하여 효과적인 수학학습지도로 연결할 수 있어야 한다(Jung et al., 2010).

순열과 조합 단원은 다른 단원에 비해 비형식적 상황의 문장제 문제가 대부분이고, 이러한 단원 특성에 따라 학생들에게 수학 불안요인을 일으킨다(Kim et al., 2009a). Kim 외 (2009a) 연구에서는 '$_nP_r=_{n-1}P_{r-1}+_{n-1}P_r$'이 성립함을 보이는 평가 문항이 해당 연구에서 가장 비형식적인 문항으로 분류하였으며, 파악해야 할 조건이 많고 간단하게 공식을 대입하여 해결할 수 없어서 학생들이 어렵게 느꼈다고 하였다. 해당 평가 문항은 2015 개정 교육과정의 고등학교 1학년 수학 교과서 9종 중 8종의 교과서에서 유사하거나 같은 문항으로 다루고 있었기 때문에 본 연구에서도 해당 문항과 동일한 형태의 문항을 GPT-4 기반의 ChatGPT의 자동 채점 가능성 탐색 문항에 포함시켰다. 또한 순열과 조합 단원이 대부분 문장제 문제이기 때문에 문제에 자연어(Natural Language)가 많이 사용되며, 답안 역시 문제의 조건 및 상황을 설명하고 서술하는 과정에서 자연어를 자주 활용할 수 있다. 자연어를 많이 활용한 답안은 같은 의미라도 다양한 방식으로 표현이 가능하기 때문에 순열과 조합 단원의 서술형 평가에서 자연어 처리에 능숙한 ChatGPT를 채점에 활용할 수 있는 가능성을 보여준다. 또한 순열과 조합 단원은 자연어가 많이 사용된 답안 이외에도 수식 위주의 답안으로 구성된 증명 문제, 자연어와 수식이 혼합되어 있는 문제 등 다양한 답안의 형태를 확인할 수 있다.

2. 서술형 평가 자동 채점

(1) 자동 채점 시스템의 원리

자동 채점 시스템은 크게 볼 때 학생 답안을 입력값으로, 답안별 점수를 출력값으로 하는 단순한 구조로 보이지만, 세부적으로는 여러 개의 알고리즘 단계가 연결된 파이프라인(Pipeline)의 구조를 가진다. Figure 1과 같이 내부 알고리즘은 일반적으로 자연어를 컴퓨터가 인식할 수 있는 형태로 변환하는 '언어 처리 단계', 변환된 형태에 채점 모델을 적용하여 점수를 부여하는 '채점 단계'로 구분할 수 있다. '언어 처리 단계'는 다시 자연어를 여러 개의 토큰으로 분리하는 토큰화(Tokenization) 단계, 분리된 토큰으로부터 채점에 불필요한 불용어(Stoword) 제거, 품사를 통일시키는 표제어 (Lemmatization) 추출 등을 실시하는 전처리(Preprocessing) 단계, 전처리된 토큰 열(Sequence)을 벡터화(Vectorization)하는 벡터화 단계 등으로 구분할 수 있다. 연구에 따라 '언어 처리 단계'는 다양하게 설계되지만, 토큰화 단계와 벡터화 단계는 언어 처리의 핵심 단계로서 거의 모든 자동 채점 시스템에서 공통적으로 거치는 단계이다.

일반적으로 토큰화 단계와 전처리 단계는 미리 구축된 형태소 사전(Dictionary)을 참조하는 방식으로 이루어지는데, 한국어 사전의 예시로는 Park 외 (2023)의 연구에서 활용한 KSS (Korean Sentence Splitter) 등이 있다. 다음으로 벡터화 단계는 연구에 따라 다양한 형태가 있을 수 있다. Noh 외 (2016)와 Park 외 (2023)는 분리된 토큰 열로부터 채점 자질(Feature)이라 불리는 특성을 추출한 후, 이를 벡터화하였다. 예를 들어, 형태소별 사용 빈도, 문장 수 등이 이에 해당한다. 하지만 이러한 방법론은 근본적으로 텍스트의 의미론적 분석 없이 에세이의 표면적 특징만을 포착한다는 한계가 있다(Chung & O'Neil Jr, 1997). 한편 분리된 토큰 열을 문서-단어 행렬(Document-Term Matrix, DTM)화하는 방법도 존재한다. DTM에서 행은 학생 답안 말뭉치(Corpus) 전체에 등장하는 토큰이다. 따라서 행벡터는 각 문서에 등장하는 토큰의 빈도에 기반한 벡터가 되는 것이다. DTM화 하는 방법 또한 토큰의 순서를 반영하지 못하는 등의 한계를 지적받을 수 있으나, 문서에 포함된 토큰의 빈도를 반영하고 있어 채점 자질을 추출하는 방법에 비해 의미를 더 반영하는 방법이라 할 수 있다. 추가로 DTM화를 개량하여 각 토큰의 중요도를 반영한 TF-IDF (Term Frequency-Inver Document Frequency) 행렬화도 널리 쓰이는 방법이다.

'채점 단계' 또한 연구에 따라 그 형태가 매우 다양하지만 채점 데이터의 필요 여부에 따라 지도 학습, 비지도 학습(Unsupervised Learning)[3]으로 구분해 볼 수 있다. 먼저 지도 학습을 활용한 채점은 전문가가 학생 답안의 일부를 미리 채점한 후, 채점 데이터를 학습시킨 채점 모델로 나머지 데이터를 채점하는 방법이다. Noh 외 (2016), Park 외 (2023)의 연구

를 비롯한 대부분의 연구가 이에 해당한다고 볼 수 있다. 지도 학습에 의한 채점은 정확도(Accuracy) 등의 모델 성능을 확보하는데 용이하다는 장점이 있다. 하지만 지도 학습 모델은 소규모 데이터로는 구축하기 어렵기 때문에 대규모 평가에 적합한 한편, 실제 학교 현장에서는 활용도가 낮다고 볼 수 있다. 또한 대규모 평가라 하더라도 학습용 채점 데이터를 생성하는 라벨링(Labeling)을 위해 많은 비용이 필요하다는 한계가 있다. 한편 비지도 학습 방법에 사용되는 학습용 답안은 정, 오답 정보가 없다. 따라서 일반적으로 채점해야 할 답안과 학습용 답안들과의 의미 유사도를 고려한 후, 가장 유사한 답안으로 분류한 후 채점하는 방식을 사용한다(Song et al., 2016). 이때, 답안과의 유사도를 계산하기 위해서는 답안을 벡터공간에 적절하게 투영해야 하는데, 고성능의 LLM에 의한 임베딩(Embedding)[4] 방법이 사용될 수 있다.

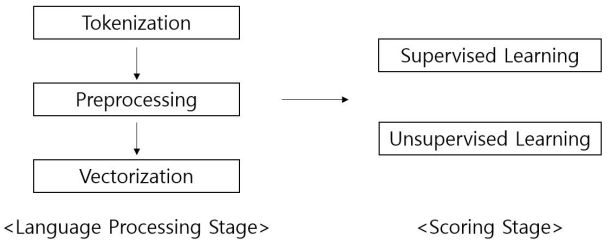

Figure 1. The automatic scoring system.

(2) 채점 모델로서 생성형 AI의 자동 채점 가능성

일반적으로 생성형 AI는 텍스트, 음성 등의 형태로 새로운 무언가를 생성하는 정교한 시스템을 의미한다(Taulli, 2023). 2018년 처음 공개된 OpenAI사의 GPT모델은 대표적인 텍스트 생성형 AI이다. GPT는 버전에 따라 구조가 약간씩 다르나, 기본적으로 사용자가 입력한 텍스트에 대해 문장 형태의 토큰 열을 출력하는 구조를 가지고 있다. 현재 시점을 기준으로 OpenAI는 최초의 GPT-1으로부터 GPT-2, GPT-3, GPT-3.5를 거쳐 GPT-4를 공개하였으며, 특히 GPT-4를 엔진으로 한 챗봇 서비스 ChatGPT는 MMLU (Massive Multitask Language Understanding) 벤치마크[5]에서 현존 언어 모델 중 최고 성능을 기록하였다(OpenAI, 2023). 뛰어난 언어 모델인 GPT-4의 등장으로 인해, 언어 모델을 활용하여 에세이를 평가하는 방법론이 대두되었다. 하지만 GPT모델의 채점자로서의 성능은 연구마다 차이가 있는 편이며, 지속적인 논의가 필요하다(Choi & Park, 2023; Yoon et al., 2023).

GPT-4 기반의 ChatGPT를 채점에 활용할 때 고려해야 할 특징은 다음과 같이 제시할 수 있다.

첫째, 텍스트의 생성이 대화 형식을 가지고 이루어진다. 이는 GPT-4 기반의 ChatGPT를 학생 답안 채점에 활용하려면 다양한 지시사항(Instruction)을 입력해야 함을 의미한다. 지시사항은 GPT-4 기반의 ChatGPT가 사용자와 의미 있는 대화를 할 수 있도록 대화의 맥락을 제공하는 서두의 텍스트를 가리킨다. 예를 들어 단순히 학생 답안만을 입력하는 경우, 학생 답안에 대한 배경 지식을 설명하는 등 채점과 무관한 텍스트를 생성한다. 따라서 학생 답안을 입력할 때마다 지시사항을 통해 GPT-4 기반의 ChatGPT에게 '당신은 교사입니다.'와 같은 역할부여, '채점하시오.'와 같은 구체적인 행동을 지시해야 한다.

둘째, '언어 처리'의 성능이 뛰어날 것으로 기대된다. '언어 처리 단계'는 자동 채점 시스템의 중요한 내부 단계로서, 자연어를 컴퓨터가 이해할 수 있는 형태로 변형하는 단계이며, 일반적으로 사전 참조 방식을 사용한다. 그러나 한국어 텍스트 처리를 위한 사전은 주로 형태소 분석에 기반하고 있기 때문에 수학 교과 지식과 수식이 포함된 텍스트에 그대로 적용하여 처리하기에는 무리가 있다. 반면 GPT-4의 구조는 밝혀지지 않았지만 GPT-3을 포함한 대부분의 트랜스포머 기반 언어 모형은 BPE (Byte Pair Encoding) 방식으로 사전을 구축하기 때문에 GPT-4 또한 그럴 것으로 추측할 수 있다. BPE 방식의 사전 구축 방법론은 형태소 분석과 달리 빈도 기반의 결정론적(Deterministic) 방법을 사용한다. 즉, 구축에 사용되는 데이터 내의 단어 등장 빈도에 따라 사전에 포함시킬 단어를 결정하기 때문에 데이터의 특성을 반영한 사전을 안정적으로 구축할 수 있다. ChatGPT의 학습 데이터의 양은 매우 거대할 뿐 아니라 MATH나 GSM-8K와 같은 수학 데이터셋 또한 다수 학습한 것으로 알려져 있어 수학 텍스트 분석을 위한 사전으로 활용해 볼 수 있다. 다음으로 ChatGPT는 영어권에서 생성하였음에도 불구하고 한국어를 포함한 26개의 언어에서 좋은 성능을 보인다는 장점과, 국가수준의 학업성취도 평가 및 대학수학능력시험의 일부 수학 문항에 대한 문제 해결 능력을 갖고 있다는 것(Kwon et al., 2023) 또한 ChatGPT를 수학 서술형 평가 채점에 활용할 수 있는 근거가

된다(OpenAI, 2023).

셋째, ChatGPT는 엔드 투 엔드 구조를 갖는다. 엔드 투 엔드 구조란 사용자의 입력값으로부터 출력값까지 한 번에 처리하는 구조를 의미한다. 앞서 언급한 바와 같이 많은 자동 채점 시스템은 다단계의 파이프라인 구조를 갖기 때문에 사용자는 각 내부 단계에 대한 높은 이해도를 갖출 필요가 있다. 하지만 ChatGPT와 같은 엔드 투 엔드 시스템은 '언어 처리 단계'와 '채점 단계'를 구분하지 않으며, 사용자가 값을 입력하기만 하면 한 번에 처리하여 출력값을 보여준다. 따라서 사용자가 AI나 딥러닝에 대한 깊은 지식을 갖추지 않아도 프롬프트만을 ChatGPT에 입력하여 모종의 결괏값을 얻을 수 있다는 장점이 있다. 하지만 한편으로는 사용자가 ChatGPT의 구조를 파악하기에는 여러 가지 한계가 있어 모형 자체를 수정하거나 보완하는 것이 어렵다는 단점이 있다.

넷째, ChatGPT를 채점 모형으로서 활용하기 위해 입력값인 '지시사항'이나 초매개변수(Hyper Parameter)' 등을 조정하며 답변의 결과를 확인하는 방법을 고려할 수 있다. 특히, 원하는 답변의 결과를 얻기 위해 지시사항 및 입력 텍스트를 조정하는 프롬프트 엔지니어링을 고려할 수 있다. ChatGPT를 채점에 활용할 경우, 일반적으로 지시사항에 채점 지침을 설정하면 각 학생 답안을 입력할 때 서두에 자동적으로 지시사항이 삽입되게 된다. 초매개변수로는 출력 텍스트의 임의성을 조정할 수 있는 온도(Temperature), 생성 텍스트의 최대 길이(Maximum Length), 토큰의 우도(Likelihood)에 따른 활용 개수를 조정할 수 있는 상위 P (Top-p) 샘플링 등이 있다. Seong과 Shin (2023)은 채점의 신뢰도 확보를 위해 Temperature를 0으로, Top-p를 1로 설정하기도 하였다.

3. 프롬프트 엔지니어링

LLM을 다운스트림 태스크(Downstream Task)[6]에 적용하기 위해서는 미세조정(Fine-Tuning)이나 프롬프트 엔지니어링을 활용할 수 있다(Lee, 2021). 미세조정은 LLM 모델에게 약간의 추가학습 데이터를 제공하여 모델의 매개변수를 업데이트하고 이를 바탕으로 의도한 목적에서 더 잘 동작하도록 만드는 것을 말한다. 하지만 미세조정 방식으로 모델 전체를 업데이트하려면 많은 비용이 든다. 한편 프롬프트 엔지니어링은 LLM 모델의 입력 프롬프트를 설계, 개선 및 최적화하는 프로세스로 사용자가 원하는 값을 출력하게 만드는 과정으로 모델을 업데이트하지 않고도 다운스트림 태스크를 바로 수행할 수 있으며(Ekin, 2023) 프롬프트 엔지니어링으로도 모델이 경쟁력 있는 태스크 수행 성능을 보이는 경우가 많다(Lee, 2021). 이에 따라 생성형 AI의 출력 결과물에 직접적인 영향을 미치는 프롬프트 엔지니어링을 위한 프롬프트 기법에 대한 연구가 활발하게 이루어지

Table 1. Prompt strategies

Prompt strategies	Description
Zero-shot	The model infers tasks without any prior examples, relying solely on its pre-existing knowledge
One-shot	The model learns from a single example or case to perform a task
Few-shot	The model learns from a small set of examples, typically ranging from 2 to 100, to grasp a task
Chain of Thought	
Zero-Shot-CoT	Inputting 'let's think step by step.' in prompts that consist of 'input, thought process, output'
Few-Shot-CoT	Based on examples that include intermediate steps in the thought process, the model also outputs final results based on intermediate steps
Self-consistency	Sampling various reasoning paths and then finding the most consistent answer among the outputted reasoning paths
Tree of thought	Implementing intermediate steps as a tree upon inputting a prompt and searching the tree to find the optimal answer
Iteratively prompt	Gradually improving and adjusting the initial prompt as a process to obtain more accurate and effective results for a specific task by optimizing the prompt
Generated knowledge prompting	Allowing the model to generate knowledge on its own, then adding the generated knowledge to the final prompt to obtain an answer
Structured prompt	Systematically organizing and clarifying the input to the artificial intelligence model to consistently obtain good results

고 있다. Table 1은 다양한 프롬프트 엔지니어링을 위한 프롬프트 기법을 나타낸 것이다.

모델이 최종 답변을 출력하기 전에 일련의 중간 단계를 생성하도록 CoT 프롬프트 기법은 복잡한 추론이 필요한 작업의 성능을 향상할 수 있다(Wei et al., 2022). 또한 Zero-Shot, One-Shot, Few-Shot 프롬프트 기법을 단독으로 사용하는 것보다 CoT 프롬프트 기법을 함께 사용하는 것이 논리적 추론 작업, 산술 추론, 기호 추론 작업에서 우수한 성능을 보였으며(Kojima et al., 2022), SC, ToT 프롬프트 기법과 같이 CoT 프롬프트 기법을 개선한 프롬프트 기법은 수학적 사고 및 계산 능력에서 더욱 월등한 성능을 보였다(Wang et al., 2022; Yao et al., 2023). 또한, 프롬프트는 점진적으로 조정하는 작업이 필요하고, 모델과 연속적인 대화 및 테스트를 통해 프롬프트를 개선할 수 있다(Wang et al., 2022). 이를 반복 프롬프트(Iteratively Prompt) 기법이라 하며 프롬프트를 최적화하기 위한 과정으로서 사용된다. 언어 모델의 경우 입력 콘텍스트(Context)의 중간에 있는 정보에 대한 정확도가 크게 저하되고, 입력 콘텍스트의 길이가 길어질수록 성능이 저하 될 수 있기 때문에(Liu et al., 2023) 반복 프롬프트 기법을 활용하여 프롬프트를 개선하는 과정에서 모델이 정확하게 파악하지 못하거나 무의미한 텍스트는 삭제할 필요가 있다.

한편 GPT 모델을 비롯한 다수의 생성형 AI는 확률 기반으로 답을 생성하기 때문에 일관되지 않은 응답이 나올 수 있어(Plevris et al., 2023) 채점의 신뢰도에 영향을 미칠 수 있다. 따라서 응답의 일관성을 높일 수 있는 SC 프롬프트 기법(Wang et al., 2022), 구조화된 프롬프트(Structured Prompt) 기법의 활용을 고려할 수 있다(Oh, 2023). SC 프롬프트 기법은 여러 번의 동일한 프롬프트를 입력하여 AI의 답변의 빈도가 높은 결과를 택하는 기법으로 산술, 상식, 논리적 추론 등을 평가하는 벤치마크 테스트에서 모델의 일관된 응답 및 추론 능력을 개선시켰다. 따라서 SC 프롬프트 기법은 답변의 일관된 응답 및 추론 능력이 필요한 서술형 평가 채점에서 효과를 발휘할 수 있을 것으로 기대할 수 있다.

구조화된 프롬프트 기법은 LLM 모델이 좋은 결과를 일관되게 얻기 위해 체계적이고 명확하게 정리하는 방식(Oh, 2023)으로 ChatGPT를 활용한 여러 교육 연구에서 활용되고 있고(Baek et al., 2023; Cho et al., 2024; Go et al., 2024; Oh, 2023; Seong & Shin, 2023), 선행 연구들은 Table 2의 프롬프트 엔지니어링 기법(OpenAI, 2024)을 근거로 구조화된 프롬프트를 작성하였다. 구조화된 프롬프트를 활용한 선행 연구들에서 ChatGPT가 수행하는 과제에 따라 프롬프트가 다르게 구성될 수 있지만 ChatGPT에게 맥락을 제공하고, 사용자가 원하는 답변의 출력 형태를 조절하도록 프롬프트를 구성한 방식은 유사하였다. Oh (2023)는 수학 문제 해결에서 효과적인 프롬프트를 고찰하는 연구를 진행했고, '역할, 규칙, 예제풀이, 문제, 과정'으로 이어지는 구조화된 프롬프트를 설계하여 9종의 고등학교 수학 교과서의 이차방정식과 이차함수 단원의 문제에 대한 ChatGPT의 정답률을 91%로 높였다. 이는 Kwon 외 (2023)의 연구에서 Zero-Shot 프롬프트 기법으로 측정했던 ChatGPT의 수학적 능력을 보다 향상시킨 결과였으며 구조화된 프롬프트의 효과를 보여주었다. Go 외 (2024)는 생성형 AI 융합 수학 수업 모형 개발 연구에서는 수식 입력이 용이한 순열과 조합 단원을 선정하고 '역할, 문제, 풀이 과정, 설명 방식'으로 프롬프트를 구조화하여 수학 문제 풀이 과정을 학습자의 수준에 맞게 설명하는 챗봇을 만들기도 하였다.

ChatGPT를 활용한 자동 채점 영역에서도 구조화된 프롬프트 기법을 적용한 사례를 찾아볼 수 있었다. Baek 외 (2023)는 과학 교과에서 GPT-3.5기반의 ChatGPT로 서술형 답안을 채점할 때 '문항, 채점 기준, 출력 규칙'으로 구조화된 프롬프트를 구성하였다. Seong과 Shin (2023)은 Baek 외 (2023)의 연구보다 최신모델인 GPT-4기반의 ChatGPT로 지리 교과의 서술형 답안을 채점하였고, 구조화된 프롬프트로 '역할, 문항, 예시 답안, 채점 기준, 출력 규칙'을 사용하였다. 하지만 두 선행연구는 문항의 특성에 따라 ChatGPT와 교사의 채점 결과 간에 유의미한 상관계수를 얻지 못하는 경우도 있었고, 부정확한 결과가 다수 출력되는 등의 한계가 있었으며 정교화된 프롬프트 개발의 필요성을 부각시켰다.

한편 수학 교과의 경우 문제 해결 전략이 2개 이상 존재(Lee et al., 2021)할 수 있고, 하나의 예시 답안을 활용하여 채점하기보다 문제 해결 전략별로 복수의 예시 답안을 작성하여 채점할 필요가 있다(Lee, 2024). 따라서 본 연구에서 선정한 3개의 문항 중 1개의 문항을 문제 해결 전략이 2개 이상 존재하는 문항으로 포함시켰고, 이를 채점하기 위해 답안을 해결 전략별로 분류하고 채점하도록 구조화된 프롬프트를 구성하였다. 또한, 해결 전략별 분류 후 채점하는 방식이 GPT-4기반의 ChatGPT의 채점 수행에 영향을 미치는지 확인하기 위해 문제 해결 전략이 1가지인 다른 2개의 문항처럼 해결 전략별로 분류하지 않고 채점하는 방법과 채점 결과를 비교하였다. 더불어 현장에서 수학 서술형 평가의 채점방식이 대부분 분석적 점수화 방법을 따르고 있음을 고려하여(Kim & Lee, 2013) 채점을 위한 프롬프트를 구성할 때 이를 반영하고자 하였다.

Table 2. Best practices for prompt engineering and six strategies for getting better results

Best practices for prompt engineering	Six strategies for getting better results
1. Use the latest model 2. Put instructions at the beginning of the prompt and use ### or """ to separate the instruction and context 3. Be specific, descriptive and as detailed as possible about the desired context, outcome, length, format, style, etc 4. Articulate the desired output format through examples 5. Start with Zero-shot, then Few-shot, neither of them worked, then fine-tune 6. Reduce "fluffy" and imprecise descriptions 7. Instead of just saying what not to do, say what to do instead 8. Code Generation Specific - Use "leading words" to nudge the model toward a particular pattern	1. Write clear instructions - Include details in your query to get more relevant answers - Ask the model to adopt a persona - Use delimiters (‘〈〉’) to clearly indicate distinct parts of the input - Specify the steps required to complete a task - Provide examples - Specify the desired length of the output 2. Provide reference text - Instruct the model to answer using a reference text - Instruct the model to answer with citations from a reference text 3. Split complex tasks into simpler subtasks - Use intent classification to identify the most relevant instructions for a user query - For dialogue applications that require very long conversations, summarize or filter previous dialogue - Summarize long documents piecewise and construct a full summary recursively 4. Give the model time to "think" - Instruct the model to work out its own solution before rushing to a conclusion - Use inner monologue or a sequence of queries to hide the model's reasoning process - Ask the model if it missed anything on previous passes 5. Use external tools - Use embeddings-based search to implement efficient knowledge retrieval - Use code execution to perform more accurate calculations or call external APIs - Give the model access to specific functions 6. Test changes systematically - Evaluate model outputs with reference to gold-standard answers

연구 방법

1. 연구 방법 및 절차

본 연구에서는 GPT-4 기반의 ChatGPT를 활용하여 서술형 채점을 진행하였으며 이하의 서술부터 GPT-4 기반의 ChatGPT를 GPT-4라고 지칭하겠다. GPT-4의 서술형 채점 성능을 확인하기 위해 학생평가지원포털(https://stas.moe.go.kr/)에 있는 고등학교 1학년 수학의 순열과 조합 문항 3개를 선정하였고, 각 문항별 예시 답안 및 채점 기준을 연구진이 GPT-4가 이해하기 쉽게 명료한 표현으로 수정하여 학생 평가 문항으로 사용하였다.

선정한 평가문항을 형성평가로 구성하여 수원에 있는 S 고등학교 1학년 8개 학급 204명의 학생을 대상으로 2023.11.21.-2023.12.05.에 한 차시당 한 문제씩 3차시에 걸쳐서 평가를 실시하였다. 문항 1, 2는 197명, 문항 3은 174명이 응답하였다. 학생들은 각 문항에 대한 서술형 답안을 종이에 풀고, 자신의 풀이 과정을 컴퓨터나 스마트폰을 활용하여 구글 설문지에 그대로 옮겨서 제출하도록 하였다. 답안에 사용되는 수식은 ‘자연수’, ‘*’, ‘!’, ‘순열기호’, ‘조합기호’ 등으로 학생들이 쉽게 입력할 수 있었다.

연구진은 학생들이 제출한 답안을 바탕으로 임의로 선정한 30개의 답안에 대해 가채점을 진행하였다. 가채점 결과 문항 1, 2는 한 가지 채점 기준으로 모든 학생들의 답안을 채점할 수 있었지만 문항 3의 경우 학생평가지원포털에 제시된 채점기준과는 상이한 해결 전략을 사용하였지만 정답으로 인정할 수 있는 한 종류의 유사 답안이 더 존재했다. 따라서 본 연구진은 문항 1, 2는 학생평가지원포털에 제시된 채점 기준을 활용한 채점 프롬프트를 구성하여 GPT-4의 채점을 진행했고, 해결 전략이 두 가지인 문항 3의 경우에는 두 가지 방식으로 채점을 진행하여 그 결과를 비교했다. 첫 번째 채점방식은 문항 3을 문항 1, 2와 같이 학생평가지원포털에 제시된 한 가지 채점 기준을 활용한 채점 프롬프트를 GPT-4에 입력하여 채점하는 방식이며, 두 번째 채점방식은 학생 답안의 유형을 분류하는 프롬프트를 GPT-4에 입력하여 답안을 분류한 뒤, 분류된 답안에 대해 각 유형별로 채점 프롬프트를 GPT-4에 입력하여 채점하는 방식이다. 두 번째 채점방식은 입력 콘텍스트의 길이가 길어질수록 성능이 저하된다는 연구(Liu et al., 2023)에 따라, Table 2의 복잡한 작업을 하위 작업으로 분할하는 프롬프트 기법을 활용한 것이다. 더불어 GPT-4의 답변의 일관성 및 추론능력을 향상시키기 위해 SC 프롬프트 기법을 활용하여 1개의 답안에 대해 GPT-4로부터 세 번씩 응답을 받아 응답의 최빈값이 있는 경우를 GPT-4의 최종 채점 결과로 판단하였다. 이상의 문항 1, 2, 3에 대한

GPT-4의 구체적인 채점 과정은 Figure 2, Figure 3과 같다.

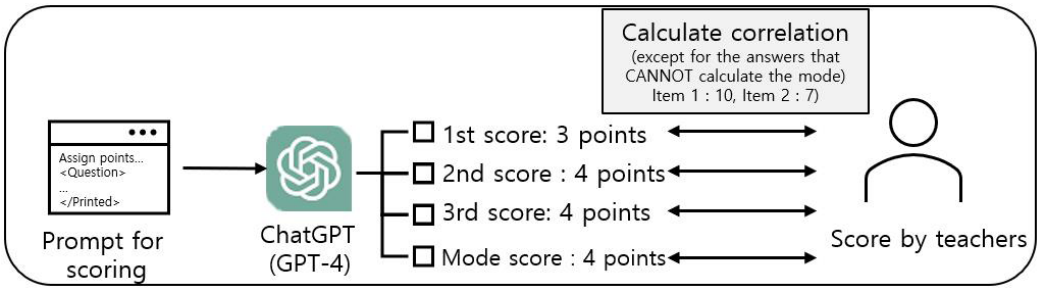

Figure 2. Scoring process for responses to items 1 and 2.

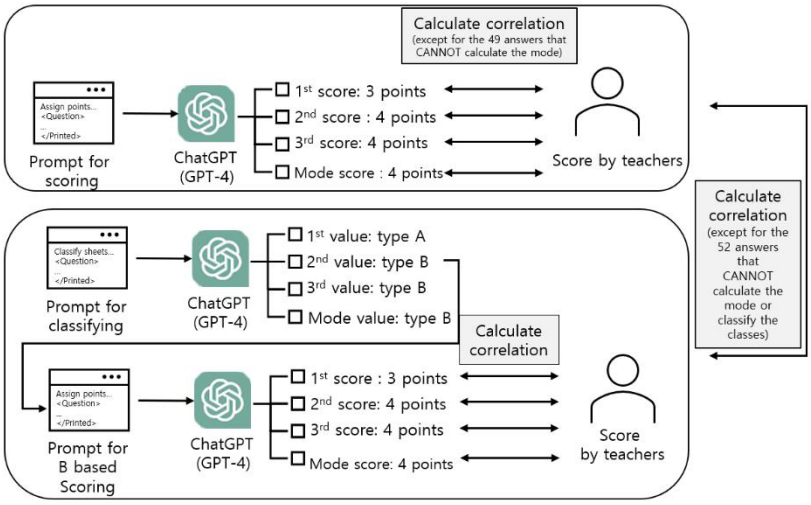

Figure 3. Classification and scoring process for the response to item 3.

다음으로 GPT-4의 채점 결과와 상관관계를 비교하기 위한 채점 데이터를 생성하였다. 이를 위해 본 연구진 중 8년 이상의 교육 경력이 있는 2명의 교사가 2023.12.20.-2024.1.15.에 문항 1, 2, 3에 대한 학생 답안을 수작업 채점하였다. 채점하는 과정에서 학생들이 제출한 문항 1, 2, 3에 대한 전체 568개 답안을 검토한 결과 구글 설문지로 학생들이 자신의 답안을 타이핑하는 과정에서 생기는 오류에 대해 일부 데이터를 전처리할 필요가 있었고, 568개의 답안을 분석한 결과 48개의 답안에서 타이핑이 미숙하여 생기는 오타가 검출되었다. 조합을 나타내는 기호 ${}_nC_r$을 'NCr, NcR'로 작성한 경우, 곱하기 기호를 문자 x로 사용한 경우 등 학생들이 답안을 작성할 때 타이핑이 아닌 손으로 작성했을 때에는 실수하지 않았을 법한 오타는 두 교사가 협의하여 오타를 수정하여 평가하였고, 비 수학적 기호 표현을 사용한 경우, 한국어 맞춤법이 틀린 경우에는 수정하지 않고 그 자체로 평가하였다.

GPT-4의 채점은 2024.01.22.-2024.01.25.에 GPT-4 API를 사용하여 구글 코랩(Google Colab) 환경에서 자동 채점을 수행하였다. 모델로부터 받은 모든 응답은 새로운 세션에서 받아 이전의 응답이 영향을 미치지 않도록 하였고, 채점 시 GPT-4의 초매개변수는 모델 초기값으로 설정하였으며, 최종 채점 프롬프트와 학생 답안을 함께 모델에 입력하여 학생 답안별 채점 점수를 획득하였다.

2. 순열과 조합 서술형 평가 문항 및 채점 기준

본 연구에 활용했던 서술형 평가 문항은 고등학교 1학년 수학 순열과 조합 단원의 학생평가지원포털에 있는 3개의 문항을 선정하여 Table 3과 같이 문제의 발문의 의미가 명료해지도록 수정하였다. 문항 1은 조합의 수 ${}_nC_r$을 두 조합의 수

의 합으로 나타내는 문제로 ${}_nC_r=\frac{n!}{r!(n-r)!}$임을 활용하여 좌변과 우변이 같음을 보여야 하기 때문에 풀이 과정에 자연어가 아닌 수식이 많이 포함된다. 문항 2는 문항 1과 같은 방식으로 조합의 수를 이용하여 증명하는 방식이 아닌 (좌변)과 (우변)의 뜻을 예로 들고 그 의미가 같다는 것을 논술하는 문제로 Kim 외 (2009a) 연구에서 가장 비형식적인 문항으로 학생들이 어려워하는 유형이었다. 문항 2의 경우 문항 1과는 달리 풀이 과정에 수식보다는 자연어로 설명하는 과정이 더 많이 포함되어 있다. 또한, 문항 1, 2는 조합에 대한 이해를 묻는 주요 문제로 2015 개정 교육과정 고등학교 1학년 수학 교과서 9종 중 문항 1은 7종, 문항 2는 8종의 교과서에 실려 있었다. 문항 3은 특정 조건을 만족하는 6자리 수의 개수를 구하는 순열 문제로 유사 문항이 9종의 모든 교과서에 나타나 있었으며 풀이 과정에 수식과 자연어가 비슷한 비중으로 사용되는 문제였다.

학생평가지원포털에서는 문항 3을 전체 경우의 수에서 여사건의 경우의 수를 제외하는 방식(Type A)에 대한 채점 기준만을 제시하였으나 유사 답안으로서 합의 법칙으로 문제를 해결한 답안을 채점하기 위한 채점 기준(Type B)을 연구진이 협의하여 추가하였다. 다음 Table 4는 문항 1, 2, 3에 대한 채점 기준이다.

Table 3. Items

Number	Items
1	Using the formula ${}_nC_r=\frac{n!}{r!(n-r)!}$, prove that the equation ${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$ holds. (Note: $1\le r<n$) [4 points]
2	The equation ${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$ implies that ${}_nC_r$, which represents the number of ways to choose r different items from n, and ${}_nC_{n-r}$ representing the number of ways to choose n−r different items from n, are equivalent. This is because choosing r different items out of n inherently means excluding the other n−r items. Thus, ${}_nC_r$ is the same as ${}_nC_{n-r}$. Illustrate this concept with examples in your discussion. (Note: $1\le r<n$) [4 points]
3	Calculate the number of numbers that can be formed by arranging the numbers 1 through 6 in a row, such that the hundred-thousand's place digit is greater than 2 and the unit's place digit is less than 5. [10 points]

Table 4. The scoring criteria for items 1, 2, 3

Item number	Scoring criteria	Score
1	Correctly expressed the term ${}_{n-1}C_{r-1}$ using a formula	1
	Correctly expressed the term ${}_{n-1}C_r$ using a formula	1
	Perfectly solved and correctly expressed the equation ${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$	2
2	When a specific individual A is designated	1
	Correctly expressed the meaning of ${}_{n-1}C_{r-1}$	1
	Correctly expressed the meaning of ${}_{n-1}C_r$	1
	Correctly expressed the meaning of the equation ${}_nC_r={}_{n-1}C_{r-1}+{}_{n-1}C_r$	1
3	Type A	
	When the total number of cases is calculated	2
	When the number of cases where the hundred-thousand's place digit is 2 or less and the cases where the unit's place digit is 5 or more are calculated separately	2
	When the number of cases where the hundred-thousand's place digit is 2 or less and the unit's place digit is 5 or more is calculated	2
	When the number of cases where the hundred-thousand's place digit is 2 or less or the unit's place digit is 5 or more is calculated	2
	When the answer is correctly calculated	2
	Type B	
	When the cases where the hundred-thousand's place digit is 3 are correctly calculated	2
	When the cases where the hundred-thousand's place digit is 4 are correctly calculated	2
	When the cases where the hundred-thousand's place digit is 5 are correctly calculated	2
	When the cases where the hundred-thousand's place digit is 6 are correctly calculated	2
	When the answer is correctly calculated using the addition principle	2

3. 채점을 위한 프롬프트 구성

GPT-4가 채점을 하기 위해 필요한 최소 맥락인 '문항, 예시 답안, 채점 기준'을 기반으로 하고, 프롬프트 기법을 활용하여 '역할, 채점 예시, 유의사항, 출력 규칙'을 추가함으로써 총 7가지 요소를 포함하는 구조화된 초기 프롬프트를 작성하였다. 서술형 평가 채점방식 중 분석적 점수화 방법에 따라 채점하기 위해 '출력 규칙' 부분에서 채점 요소별로 출력할 수 있게 구성하였다. 이는 사고 과정의 중간 단계를 기반으로 최종결과를 출력하는 CoT 프롬프트 기법을 사용한 것으로 GPT-4가 채점을 수행할 때, 점수만 출력하게 하는 것이 아니라 각 항목에 대한 점수를 출력하기 직전에 점수에 대한 근거를 구체적으로 서술하도록 하여 GPT-4의 추론 능력을 향상시키기 위함이다. 실제로 Seong과 Shin (2023)의 연구에서도 AI의 분석적 채점 방법을 사용한 채점 결과가 총체적 채점 방법을 사용한 것보다 교사 채점과의 상관계수가 더 높았다. '채점 예시'는 교사의 가채점 결과 모두 다른 점수를 받은 4개의 학생 답안을 임의로 선정한 뒤 교사의 채점 결과를 '출력 규칙' 양식과 동일하게 작성하여 입력했다. 즉, Few-Shot으로 제공한 채점 예시에서도 학생 답안의 채점 근거를 먼저 서술하고 해당 영역의 점수를 뒤에 부여하여 GPT-4가 점수를 출력하기 전에 앞의 내용을 기반으로 점수를 논리적으로 추론할 수 있도록 유도하였다.

한편 GPT-4는 입력된 프롬프트의 맨 앞, 맨 뒤의 정보에 대한 기억이 가장 정확하다는 Liu 외 (2023)의 연구에 따라 Table 2의 프롬프트 기법에 따라 맨 앞에는 마크다운 언어에서 사용하는 헤더 표현인 '###'를 활용하여 '역할'을 부여하였고, 일정한 형태로의 출력을 위해 '출력 규칙'을 맨 뒤에 배치하였다. GPT-4가 0.5 단위의 소수점 점수를 부여하는 경향이 있어 Html의 문법에서 주석을 나타내는 기호인 〈!-- --〉를 사용하여 '〈!-- 0.5, 1.5점 부여 시 반올림한다.--〉', '배점은 0점 또는 1점 또는 2점을 부여할 수 있다.' 등의 표현을 '유의사항'에 추가하여 프롬프트를 점진적으로 개선해나갔다. 또한 Table 2에 따라 프롬프트의 문단은 '〈〉, 〈/〉'를 이용하여 구분하였으며 문항 1, 2에 대한 최종 채점 프롬프트는 Table 5와 같다.

Table 5. The scoring prompt in item 1 and 2

	Scoring prompt for scoring item 1	Scoring prompt for scoring item 2
Role	### You are Scoring a mathematical descriptive answer. Refer to the 〈Example Answer〉, 〈Scoring Criteria〉, 〈Important Notes〉, and 〈Scoring Examples〉 for the following 〈Item〉, and output the results according to the 〈Output rules.〉 ###	
Item/Example answer/Scoring criteria		
Important notes	〈Important Notes〉 1. Scoring Criteria 1 and 2 can be awarded 0 or 1 point. 2. Scoring Criteria 3 can be awarded 0, 1, or 2 points. 3. Grade by referring to the 〈Scoring Examples〉. 4. Adhere to the 〈Output〉 rules. 5. If non-mathematical symbols are used, deduct 1 point from the total score. 6. There must be 'formula calculation'. If only the meanings of the left and right sides are simply described, no points are awarded. 〈/Important Notes〉	〈Important Notes〉 1. Scoring Criteria 1, 2, 3, and 4 can be awarded 0 or 1 point. 2. Grade by referring to the 〈Scoring Examples〉. 3. Adhere to the 〈Output〉 rules. 4. If non-mathematical symbols are used, deduct 1 point from the total score. 5. No points are awarded if the left and right sides are proven to be equal through 'formula calculation'. 〈/Important Notes〉
Scoring examples (provide four examples graded by the teacher as a few-shot demonstration)		
Output rules	〈Output Rules〉〈!--Write the content about the 〈Student Answer〉 in the section enclosed by '[ ]'. --〉 1. For correctly expressing n-1Cr-1 as an equation: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 2. For correctly expressing n-1Cr as an equation: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 3. For nCr=n-1Cr-1 + n-1Cr: [Scoring Rationale] [0 or 1 or 2] points 〈!-- Round up if 0.5 or 1.5 points are given.--〉 4. Total Score [ ] points 〈/Output Rules〉	〈Output Rules〉〈!--Write the content about the 〈Student Answer〉 in the section enclosed by '[ ]'. --〉 1. If a specific individual A is set: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 2. For correctly expressing the meaning of n-1Cr-1: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 3. For correctly expressing the meaning of n-1Cr: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 4. For correctly expressing the meaning of nCr=n-1Cr-1 + n-1Cr: [Scoring Rationale] [0 or 1] point 〈!-- Round up if 0.5 points are given. --〉 5. Total Score [ ] points 〈/Output Rules〉

Table 6. Item 3 answer type classification prompt

	Item 3 answer type classification prompt
Role	### You are to classify a math 〈Student Answer〉 provided by the User as either '〈Type A〉' or '〈Type B〉', based on the approach to the item-solving process, and output according to the 〈Output Rules〉. Note that classification is based on the approach taken in the solution process, and errors in the calculation process of the provided answers are not considered. ###
Item 3	
Type	〈Type A〉 〈!--Approach: Type that resolves the item by 'subtracting the number of cases in the complement from the total number of cases'--〉 The number of ways to arrange 6 numbers in a row is 6!=6*5*4*3*2*1=720. The number of cases where the digit in the hundred-thousands place is 2 or less is 2*5!=2*5*4*3*2*1=240. The number of cases where the digit in the ones place is 5 or more is 2*5!=2*5*4*3*2*1=240. The number of cases where the digit in the hundred-thousands place is 2 or less, and the digit in the ones place is 5 or more is 2*2*4!=2*2*4*3*2*1=96. Therefore, the number of cases where the hundred-thousands digit is 2 or less, or the ones digit is 5 or more is 240+240-96=384. Thus, the number of cases we are looking for is 720-384=336. 〈/Type A〉 〈Type B〉 〈!--Approach: Type that solves the item by 'dividing the cases according to a certain criterion and summing each case' using the addition principle--〉 i) If the digit in the hundred-thousands place is 3: there are 3 possible digits for the ones place, which are 1, 2, 4, so there are 3 cases, and the number of ways to arrange the other digits is 4!, so 3*4!=72. ii) If the digit in the hundred-thousands place is 4: there are 3 possible digits for the ones place, which are 1, 2, 3, so there are 3 cases, and the number of ways to arrange the other digits is 4!, so 3*4!=72. iii) If the digit in the hundred-thousands place is 5: there are 4 possible digits for the ones place, which are 1, 2, 3, 4, so there are 4 cases, and the number of ways to arrange the other digits is 4!, so 4*4!=96. iv) If the digit in the hundred-thousands place is 6: there are 4 possible digits for the ones place, which are 1, 2, 3, 4, so there are 4 cases, and the number of ways to arrange the other digits is 4!, so 4*4!=96. Therefore, the total number of cases we are looking for is 72+72+96+96=336. 〈/Type B〉
Output rules	〈Output Rules〉 〈!--Review the student answer to see if it is similar to Type A or Type B.--〉 1. [〈!--Review the student answer--〉] 2. [〈!--Evaluate the similarity of the student answer to Type A. Focus on the approach of the student answer, and do not consider calculation errors in the similarity classification.--〉] 3. [〈!--Evaluate the similarity of the student answer to Type B. Focus on the approach of the student answer, and do not consider calculation errors in the similarity classification.--〉] 4. [〈!--Describe the basis for which type, either Type A or Type B, the student answer's approach is similar to--〉] 5. [〈!--Output the result whether it is similar to Type A or B.--〉 'This answer is similar to Type [].']〈/Output Rules〉

Table 7. The scoring prompt by answer type for item 3

	Type A scoring prompt	Type B scoring prompt
Role	### You are Scoring a mathematical descriptive answer. Refer to the 〈Example Answer〉, 〈Scoring Criteria〉, 〈Important Notes〉, and 〈Scoring Examples〉 for the following 〈Item〉, and output the results according to the 〈Output rules.〉 ###	
Item 3/Example answer/Scoring criteria		
Important notes	〈ImportantNotes〉1.Scoring Criteria 1, 4, 5, 6 can be awarded 0, 1, or 2 points. 2. Scoring Criteria 2, 3 can be awarded 0 or 1 point. 3. Grade by referring to the 〈ScoringExamples〉. 4. Adhere to the 〈Output〉 rules. 5. Deduct 1 point from the total score if non-mathematical symbols such as '-〉, =〉' are used.〈/Important Notes〉	
Scoring examples (provide four examples graded by the teacher as a few-shot demonstration)		
Output rules	〈Output Rules〉〈!--Write the content regarding the 〈Student Answer〉 in the section enclosed by '[ ]'. --〉 1. If the total number of cases is calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 2. If the number of cases where the digit in the hundred-thousands place is 2 or less is calculated: [Scoring Rationale] [0 or 1 point] 〈!-- Round up if 0.5 points are given. --〉 3. If the number of cases where the digit in the ones place is 5 or more is calculated: [Scoring Rationale] [0 or 1 point] 〈!-- Round up if 0.5 points are given. --〉 4. If the number of cases where the digit in the hundred-thousands place is 2 or less and the digit in the ones place is 5 or more is calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 5. If the number of cases where the digit in the hundred-thousands place is 2 or less or the digit in the ones place is 5 or more is calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 6. If the answer is correctly calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 7. Total Score [ ] points 〈/Output Rules〉	〈Output Rules〉〈!--Write the content regarding the 〈StudentAnswer〉in the section enclosed by '[ ]'. --〉 1. If the cases where the digit in the hundred-thousands place is 3 are accurately calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 2. If the cases where the digit in the hundred-thousands place is 4 are accurately calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 3. If the cases where the digit in the hundred-thousands place is 5 are accurately calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 4. If the cases where the digit in the hundred-thousands place is 6 are accurately calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 5. If the answer is correctly calculated: [Scoring Rationale] [0, 1, or 2 points] 〈!-- Round up if 0.5 or 1.5 points are given.--〉 6. Total Score [ ] points 〈/Output Rules〉

한편 문항 3은 Figure 3에서 언급한 것처럼 두 가지 방식으로 프롬프트를 구성하였다. 먼저 학생평가 지원포털에 나온 채점 기준을 근거로 문항 1, 2와 같이 7가지로 구성된 구조화된 프롬프트를 구성하였다. 다음으로 Table 6과 같이 '역할, 문항, 유형, 출력 규칙' 4가지 요소로 답안의 문제 해결 전략에 따른 유형을 분류하는 프롬프트를 만들고, Table 7과 같이 '역할, 문항, 예시 답안, 채점 기준, 유의사항, 채점 예시, 출력 규칙' 7가지 요소로 각 유형에 맞는 채점 프롬프트를 구성하였다. 즉, 학생 답안이 여사건을 활용한 방식의 풀이인지, 합의 법칙을 활용한 방식의 풀이인지를 분류하는 프롬프트와 각 유형별 채점 프롬프트를 각각 작성하였다.

결과 분석 및 논의

1. 문항 1, 2에 대한 채점 결과

두 명의 교사가 문항 1, 2에 대한 채점을 진행할 때 학생평가지원포털에 제시된 분석적 점수화 방법의 채점 기준을 바탕으로 임의로 선택한 30개의 문항에 대해 가채점을 진행하였고, 각각 1회씩 채점한 뒤 1회씩 교차 검토하였다. 채점 과정에서 답안의 서술이 모호한 것은 필요한 부분은 동일 장소에서 즉각적으로 충분한 협의를 거쳤으며 문항 1, 2에 대한 채점 기준이 명확하고 협의가 잘 진행되어 1회 채점 및 교차 검토 이후 2차 채점이 필요할 만큼의 큰 이견은 없었다.

문항 1은 두 교사와 GPT-4 모두 총 197개의 답안을 채점하였다. 생성형 AI는 확률 기반으로 답변을 생성하여 매 출력마다 결과물이 달라질 수 있기 때문에 보다 일관적인 결과를 얻기 위해 SC 프롬프트 기법을 적용하였다. 따라서 세 번의 GPT-4 채점 결과에서 최빈값을 산출할 수 있는 답안을 대상으로 상관관계 분석을 진행하였고, 최빈값이 나타나지 않았던 10개의 데이터(세 개의 채점 결과가 모두 다름)를 제외한 187개의 답안 데이터가 분석 대상이 되었다. Table 5에 나타나 있는 프롬프트를 GPT-4에 세 번씩 입력하여 얻은 채점 결과 및 채점 결과의 최빈값과 두 교사의 채점 결과 사이의 상관계수는 Table 8과 같다.

분석 결과 문항 1에 대한 두 교사 채점 점수의 상관계수는 0.964로 채점자 간의 일치도가 매우 높았고, 이는 각 교사와 GPT-4의 개별 채점 점수의 상관계수보다도 높았다. 또한 교사 B와 GPT-4의 세 번째 채점 점수와의 상관계수를 제외하고 모두 0.8 이상의 값을 가져 교사의 채점과 GPT-4의 채점과의 강한 상관이 있다고 판단할 수 있었다. 이러한 결과는 Few-Shot-CoT, 구조화된 프롬프트 기법을 활용하여 구성한 프롬프트가 채점에서 교사 채점자와 어느 정도 유사하게 채점하고 있음을 보여준다. 각 교사의 채점 점수와 GPT-4의 세 번의 채점 점수의 최빈값 사이의 상관계수는 0.872, 0.905로 개별 GPT-4 채점 점수와의 상관계수보다 높아 SC 프롬프트 기법을 통해 교사 채점자와 더욱 유사한 결과를 낼 수 있음을 시사한다. 즉, GPT-4의 특성상 같은 프롬프트를 여러 번 입력했을 때 매번 다른 결과가 나올 수 있는데 SC 프롬프트 기법은 개별 GPT-4들의 채점 결과값의 최빈값을 최종 채점 결과값으로 정하기 때문에 GPT-4의 일관적인 채점 결과를 얻음과 동시에 교사의 채점 결과값과 더욱 정합성 있게 접근할 수 있다.

Table 8. Analysis of correlation for item 1 scoring

	Teacher A	Teacher B	GPT-4 1st	GPT-4 2nd	GPT-4 3rd	Mode
Teacher A	1	0.964	0.891	0.873	0.802	0.905
Teacher B	–	1	0.845	0.860	0.792	0.872

문항 2의 경우도 두 교사와 GPT-4 모두 197개의 답안을 채점하였다. 세 번의 GPT-4 채점 결과에서 최빈값이 산출되지 않는 답안(세 개의 채점 결과가 모두 다름)이 7개였으며, 이를 제외한 190개 답안에 대해 상관관계를 분석하였다. 먼저 두 교사 채점 결과 간 상관계수는 0.958로, 문항 1과 마찬가지로 문항 2도 교사 채점 점수간 상관관계가 상당히 높았다. 교사 채점 점수와 GPT-4 채점 점수 간의 상관계수는 0.795-0.854 정도로, 교사 채점 점수간 상관계수보다 낮지만 비교적 높은 수준으로 형성되었으며 SC 프롬프트 기법에 의해 산출한 GPT-4 채점 점수의 최빈값과 교사 채점 점수간 상관계수는 0.834, 0.861로 소폭 상승하였다. 문항 2의 채점에 대한 상관관계 분석 결과를 정리하면 Table 9와 같다.

Table 9. Analysis of correlation for item 2 scoring

	Teacher A	Teacher B	GPT-4 1st	GPT-4 2nd	GPT-4 3rd	Mode
Teacher A	1	0.958	0.817	0.835	0.854	0.861
Teacher B	–	1	0.795	0.804	0.832	0.834

2. 문항 3에 대한 채점 결과

두 명의 교사가 문항 3에 대한 채점을 진행할 때 문항 1, 2와 동일한 방식으로 학생평가지원포털에 제시된 분석적 점수화 방법의 채점 기준을 바탕으로 임의로 선택한 30개의 문항에 대해 가채점을 진행하였다. 가채점 결과 학생평가지원포털에 제시된 '여사건을 활용한 답안' 외에도 '합의 법칙을 활용한 답안' 역시 정답으로 인정할 수 있다고 판단하였다. 이에 따른 추가적인 채점 기준이 필요하다고 판단하여 Table 4처럼 '합의 법칙을 활용한 답안'의 채점 기준을 추가하였다. 두 명의 교사는 문항 3에 대한 답안을 각각 1회씩 채점한 뒤 1회씩 교차 검토하였고, 채점 과정에서 답안의 서술이 모호한 것은 필요한 부분은 동일 장소에서 즉각적으로 충분한 협의를 거쳤다. 문항 1, 2와 동일하게 문항 3 또한 협의가 잘 진행되어 1회 채점 및 교차 검토 이후 2차 채점이 필요할 만큼의 큰 이견은 없었다.

문제 해결 전략이 2가지인 문항 3의 경우 두 교사와 GPT-4 모두 174개의 답안을 채점했고, '문항 1, 2와 같은 방법으로 채점하는 방식'과 '해결 전략별로 답안을 분류 후 채점하는 방식'으로 채점을 진행하였다. 첫번째 방식으로 채점을 수행했을 때 최빈값이 나타나지 않는 데이터가 49개였으며 이 데이터를 제외한 125개의 데이터에서 상관관계를 분석하였다. 두 교사 채점 결과 간 상관계수는 0.965로 매우 강한 상관관계가 있었고, 교사 채점 점수와 각 GPT-4 채점 점수의 상관계수는 문항 1, 2에서 나타났던 상관계수보다 낮은 값인 0.740-0.794 수준으로 두 값이 상관이 있다고 해석할 수 있었다. 문항 3의 125개의 데이터에 대한 상관관계 분석 결과를 정리하면 Table 10과 같다.

Table 10. Analysis of correlation for item 3 scoring

	Teacher A	Teacher B	GPT-4 1st	GPT-4 2nd	GPT-4 3rd	Mode
Teacher A	1	0.965	0.756	0.794	0.784	0.769
Teacher B		1	0.740	0.793	0.769	0.758

문항 3은 문항 1, 2와는 다르게 하나의 채점 기준으로는 정확하게 채점하기 어려운 유사 답안의 종류가 2가지였기 때문에 답안을 해당 채점 기준으로 해석하는 과정에서 GPT-4가 어려움을 겪은 것으로 판단할 수 있다. 따라서 해결 방법이 여러 가지인 답안을 GPT-4를 활용하여 채점하는 경우 해결 전략별로 채점기준을 수립할 필요가 있다고 판단할 수 있다.

다음은 분류 프롬프트를 넣은 GPT-4를 활용하여 해결 전략별로 답안을 분류한 뒤 해결 전략별 채점 프롬프트를 넣은 GPT-4를 활용하여 채점한 결과이다. Table 11과 같이 174개의 답안을 Type A (여사건의 방식), Type B (합의 법칙) 또는 미분류로 분류한 결과, 두 교사의 분류 결과는 100% 일치하였고, GPT-4도 교사와 약 94%-96% 일치하였으며 최빈값이 나오지 않은 데이터는 단 한 개가 있었다. 이를 포함하여 GPT-4는 7-11개 답안에 대해 교사와 다르게 분류하였는데 해당 답안들과 최빈값이 나오지 않은 답안을 검토한 결과 교사들이 Type B 또는 미분류로 분류한 답안이었고, 이 답안들의 특징은 설명 없이 답만 작성되어 있거나 곱의 법칙으로만 답이 서술되어 있었다.

Table 11. Consistency rate for item classification

	Teacher B	GPT-4 1st	GPT-4 2nd	GPT-4 3rd	Mode
Teacher A	100 (%)	96.000 (%)	93.714 (%)	93.714 (%)	96.000 (%)

교사는 곱의 법칙으로만 작성한 답안을 식 안에 들어있는 합의 법칙 요소를 파악하여 합의 법칙을 사용한 Type B로 분류하였다면 GPT-4는 이를 미분류로 분류하거나 세 번의 분류 시도가 모두 달라 최빈값이 나타나지 않는 특징을 보였다. Table 12

의 첫 번째 학생 답안은 오답에 해당하는 답안이지만 '4*4*3*2*1*4=384'를 계산하는 과정 속에는 십만 자리에 올 수 있는 수 4가지의 경우마다 '4*3*2*1*4' 가지의 경우가 생긴다고 해석하는 것에서 합의 법칙을 활용했다고 볼 수 있다. 또한 두 번째 학생 답안 역시 십만의 자리에 올 수 있는 숫자 4가지를 기준으로 나머지 5개의 자리에 올 수 있는 경우가 모두 같다고 해석하여 이를 곱하여 계산하는 오류를 범했지만 이는 각 경우를 동일하게 4번을 더한 것과 같은 방식으로 합의 법칙을 활용한 풀이 과정이라고 볼 수 있다.

Table 12. Student answer

	Student answer
1	Among the numbers from 1 to 6, there are 4 numbers greater than 2, which are 3, 4, 5, and 6, thus for the hundred-thousand's place, there are 4 options (4C1, which means 4 choices). Similarly, for numbers less than 5, we have 4, 3, 2, and 1, meaning the unit's place also has 4 options (4C1, 4 choices). The remaining 4 digits have no restrictions, so we arrange these 4 digits excluding the hundred-thousand's and unit's places. Since all these processes occur simultaneously, the total is 4*4*3*2*1*4=384.
2	Hundred-thousand's place numbers that cannot use 1,2=4 options. Unit's place numbers that cannot use 5, 6 and must be different from the hundred-thousand's place number=3 options. (6 total numbers)-(number of options for the hundred-thousand's and unit's places) 4 options. Repeat the same method 2 more times. 4*3*4*3*2=288.

문항 3에서 GPT-4는 Type A의 학생 답안은 100% 정확도로 분류해 냈지만 Type B의 경우 6개의 학생 답안을 Type B로 분류하지 못했다. 이는 토큰의 등장 형태 때문인데 Type A는 '-' 기호, 전체 경우의 수인 '720', '빼면'과 같이 단일 토큰만으로도 분류해 낼 수 있는 반면, Type B는 '십만 자리'를 이용한 방법, '일의 자리'를 이용한 방법, '직접 세기'를 이용한 방법 등 다양한 토큰들이 사용되어 일련의 논리 단계를 거쳐야만 합의 법칙을 사용하고 있음을 인식할 수 있다는 것이다. 따라서 이러한 결과는 GPT-4가 프롬프트에 입력된 텍스트 내에서의 분석에 비해 일련의 논리 단계를 거친 텍스트 분석에는 다소 성능이 떨어질 수 있음을 의미한다고 볼 수 있다.

한편 GPT-4가 분류한 전체 174개의 답안에서 최빈값이 나타나지 않는 1개의 문항을 제외한 173개의 답안 중 미분류 데이터가 7개였고, 이를 제외한 166개의 답안을 여사건 방법 또는 합의 법칙을 활용한 방법으로 채점하였다. 166개의 답안을 대상으로 GPT-4로 세 번의 채점 결과 17개의 답안에서 최빈값이 나오지 않았으며 28개의 답안을 해당 채점 기준으로 채점이 불가능하다고 응답하였다. 28개의 답안은 모두 Type B인 합의 법칙을 활용하여 작성한 답안이었으며 분석 결과 예시 답안으로 주었던 십만의 자리를 기준으로 합의 법칙을 사용한 답안이 아닌 일의 자리를 기준으로 나누어 서술한 답안, 십만의 자리와 일의 자리를 동시에 고려하여 서술한 답안, 서술 과정이 비 논리적인 답안 등이 있었다. 28개 중 6개를 제외한 22개 문항은 두 교사 모두 0점을 부여한 답안으로 GPT-4가 채점이 불가능하다고 응답한 대부분의 답안이 서술형 채점에서 점수를 받을 수 없는 답안이었다. 따라서 166개의 답안 중 최빈값이 나오지 않은 17개의 답안과 해당 채점 기준으로 채점이 불가능한 답안 28개의 합인 총 45개를 제외한 121개의 답안에 대한 두 교사와 GPT-4의 상관관계 분석 결과를 정리하면 Table 13과 같다.

Table 13. Analysis of correlation for item 3 scoring

	Teacher A	Teacher B	GPT-4 1st	GPT-4 2nd	GPT-4 3rd	Mode
Teacher A	1	0.957	0.920	0.911	0.898	0.931
Teacher B		1	0.917	0.908	0.896	0.928

121개 문항에 대한 두 교사 채점 결과 간 상관계수는 0.957로 매우 강한 상관관계가 있었고, 교사 채점 점수와 각 GPT-4 채점 점수의 상관계수는 0.896-0.920 수준으로 문항 1, 2에서의 결과 및 분류를 하지 않고 채점했던 문항 3에서의 결과보다 높았으며 교사 채점 점수와 최빈값에 대한 상관계수는 0.928, 0.931로 매우 높았다. 이는 한 문항을 해결할 수 있는 문제 해결 전략이 여러 개인 문항의 경우 답안을 해결 전략에 따라 분류한 뒤 각 채점 기준에 맞는 프롬프트를 활용하여 채점한다면 GPT-4가 교사의 채점과 더욱 유사하게 기능하게 할 수 있는 가능성을 시사한다.

결론 및 제언

본 연구에서는 수학 교과 서술형 평가 자동 채점에 GPT-4를 활용하고 그 결과를 분석하였다. 특히, 채점 모형으로서 GPT-4를 활용하기 위하여 Few-Shot CoT, SC, 구조화된 프롬프트 기법 등을 적용한 채점 방법론을 제안하였으며, 실제 학생 답안 데이터에 이를 적용한 후 상관관계를 분석하였다. 상관관계 분석 결과에 따라 내릴 수 있는 결론은 다음과 같다.

첫째, 문항 1, 2, 3에서 공통적으로 GPT-4의 채점 점수와 교사 채점 점수 간의 상관계수는 약 0.8 수준으로, 둘 사이에 강한 상관관계가 있음을 나타냈다. 이는 GPT-4의 채점이 교사의 채점과 유사한 수준에서 이루어질 수 있음을 의미하며, 교사 채점자가 GPT-4를 채점의 신뢰도 확보 등을 목적으로 한 보조 수단으로 활용할 수 있는 가능성을 보여준다.

둘째, 문항 1, 2, 3에서 GPT-4가 한 문항당 세 번씩 채점한 점수의 최빈값과 교사와의 상관계수가 단일 채점 상관계수 값보다 모두 소폭 상승하였다. 이는 최빈값을 활용한 SC 프롬프트 기법이 GPT-4를 활용한 채점의 신뢰도를 높일 수 있는 가능성을 나타낸 것이라 할 수 있다.

셋째, 문제 해결 전략이 다양한 문제에 대해 프롬프트 엔지니어링을 통해 답안을 '분류' 하고, 분류된 문항을 해당 채점 기준에 맞게 채점하면 GPT-4의 채점 성능을 향상할 수 있음을 확인하였다. 특히, GPT-4의 3번의 분류의 결과에 대한 최빈값은 174개의 답안 중 7개만 교사와 달랐기 때문에 분류 성능은 매우 우수하다는 것을 확인할 수 있었다. 문항 3에서 답안의 유형 분류 및 유형별 채점 점수와 교사 채점 점수간 상관계수는 문항 1, 2에서와 같이 답안을 분류하지 않고 채점한 GPT-4의 채점 점수와 교사 채점 점수간 상관계수보다 상당히 높은 수준으로 산출되었다. 이는 여러 해결 전략이 존재하는 문항에 대하여 교사가 문항 분석, 가채점 등을 통해 이를 인식하고 해결 전략별로 학생 답안을 분류하는 프롬프트를 입력한다면 채점의 신뢰도를 높일 수 있음을 나타낸 것이다.

넷째, 답안에 수식이 많이 포함되어 있는 문항 1, 수식보다 자연어가 많이 포함되어 있는 문항 2, 자연어와 수식이 고르게 섞여있는 문항 3에서 다양한 프롬프트 기법을 활용하여 채점을 진행하였다. GPT-4는 일반적으로 자연어 처리에도 능숙하지만 순열과 조합 단원에서는 수식 처리도 우수함을 확인할 수 있었다. Kwon 외 (2023), Oh (2023)의 연구에서 대학수학능력시험, 학업성취도평가, 수학 교과서에 있는 문항에 대한 ChatGPT의 수학적 능력을 파악하기 위해 '∑, log, ∫, lim' 등 수식을 Equatio 등의 프로그램을 활용하여 LaTeX 문법으로 변환하여 입력하였지만 본 연구에서는 수식 자체 그대로 입력했고, 이를 통해 GPT-4가 LaTex 문법으로 표현하지 않은 수식도 잘 처리함을 확인할 수 있었다. 즉, '$_nC_r=_{n-1}C_{r-1}+_{n-1}C_r$'와 같은 수식을 GPT-4에 LaTex 문법으로 변환하여 입력하지 않아도 구조화된 프롬프트 속에 있는 맥락을 바탕으로 조합에 관한 식으로 정확하게 인식하여 채점하였다.

한편 본 연구의 한계는 다음과 같다.

첫째, 본 연구는 OpenAI의 GPT-4의 API를 사용하여 구글 코랩에서 학생들의 답안을 채점하였다. GPT 모델은 지속적으로 업데이트되고 있고, GPT 이외의 모델이나 앞으로 나올 상위모델이 본 연구에서 사용했던 프롬프트가 연구결과와 비슷한 수준으로 작동할지는 알 수 없다. 즉, 모델의 업데이트 및 모델의 종류에 따라서 최적화된 시스템 프롬프트 디자인을 수립해야 하기 때문에 이에 대한 연구가 필요하다.

둘째, GPT-4 API는 입력 및 출력 토큰에 따라 비용을 계산한다. 본 연구에서 3개의 문항을 채점하면서 지불한 금액은 약 450달러였다. 최빈값을 활용한 SC 프롬프트 기법은 같은 질문을 3회 반복하여 모델의 최종적인 채점 결과를 결정하는 방식에서 모델의 사용량이 많아 비용이 높게 발생되었다. 이는 단위 학교에서 시도하기에는 다소 높은 금액으로, 현장에서의 활용성을 낮추는 원인이 될 수 있고, 데이터를 직접 이용한 지도 학습 기반 자동 채점 연구의 높은 비용 문제를 극복하지 못했다는 한계를 보여준다.

셋째, 본 연구는 GPT-4의 채점 성능 측정을 위한 지표로서 교사가 부여한 채점 점수와의 상관관계를 비교하는 수준에 그치고 있다. 이는 GPT-4가 교사 채점자와 전반적으로 유사한 점수를 부여할 수 있음을 의미하지만 GPT-4의 채점 성능 자체가 높음을 의미하는 것은 아니다. 게다가 단 한 명의 학생 답안에 대해서도 잘못된 채점이 이루어지지 않아야 하는 교육 평가 영역에서는 교사의 개입을 배제하는 것은 다소 무리가 있어 보인다.

본 연구의 제언은 다음과 같다.

첫째, 본 연구에서는 모델의 추가적인 학습이나 내부 파라미터를 조절하지 않고, 프롬프트 엔지니어링에 주목하여 서술형 자동 채점 연구를 수행하였다. 따라서 프롬프트 엔지니어링 외의 방법인 파인튜닝을 활용한 채점 수행의 가능성에 대한 연구도 고려해볼 수 있다.

둘째, 채점의 핵심 과정을 GPT-4에 전적으로 의존하면 OpenAI의 과금 정책에 따라 큰 비용이 발생할 수 있고, 해외 기업

인 OpenAI에 수학 문항, 학생 답안, 채점 기준 등의 데이터를 제공하는 것에 대한 우려가 발생 할 수 있다. 따라서 우리나라의 독자적 채점 모델 개발 등의 방법을 통해 이러한 한계를 보완할 필요가 있다.

셋째, 본 연구에서는 GPT-4의 채점 근거를 바탕으로 정확한 채점 수행 여부의 가능성에 초점을 맞추었지만 서술형 자동 평가의 최종적인 목적은 학생 개별 피드백 제공이 되어야 한다. 단순히 정답 및 오답 여부에 따른 평가, 학생들이 작성한 서술형 답안에 대한 총체적 피드백 뿐만 아니라 서술하는 과정에서의 국소적인 모든 부분에 대한 피드백을 제공하는 것에 목표를 두어야 한다.

넷째, 본 연구에서는 학생 답안을 직접 타이핑하여 데이터화 하였다. 하지만 수식을 입력하기에 타이핑은 적합하지 않으며, 타이핑 과정에서 오탈자가 발생하기도 하였다. 따라서 OCR 기법 등을 활용하여 한글 손글씨와 수식을 손쉽게 데이터화하는 방법이나 수식 편집기를 활용하여 학생들이 수식이 포함된 답안을 쉽게 입력할 수 있도록 하는 방법 등이 필요하다. 나아가 이러한 방법은 단위 학교에서 매해 폐기되고 있는 수행평가, 형성평가 등의 서술형 답안을 데이터화하여 향후 자동 채점 모델을 만드는 데 기여할 수 있다.

다섯째, 본 연구에서는 수학 영역에서 비교적 자연어가 많이 사용되는 순열과 조합 단원에 대하여 GPT-4의 채점을 탐색하였다. 따라서 수식 사용 비중이 높은 '대수', '해석' 영역의 문항에 대한 채점, '기하' 영역과 같이 그림에 대한 이해를 바탕으로 문제를 해결해야 하는 문항의 채점 수행능력도 확인해 볼 필요가 있다.

여섯째, 문제 해결 전략이 2가지 이상인 문항을 자동 채점할 때 학생들의 모든 답안을 채점하기 전까지는 학생 답안에서 나타나는 문제의 해결 전략의 총 개수는 알 수 없다. 하지만 서술형 평가를 실시하는 교사가 미리 예상한 문제 해결 전략에 부합하는 채점 프롬프트를 구성하고, 전체 학생 답안에서 해당 문제 해결 전략을 사용한 학생들의 답안을 '추출'하여 채점하는 것은 충분히 가능하다. 본 연구에서 GPT-4 는 답안을 문제 해결 전략별로 '분류' 하는 것에서 교사와 96% 일치했기 때문에 '추출'하여 채점하는 것도 교사와 비슷한 수준의 성능을 보일 것으로 예상할 수 있다. 따라서 대규모 평가에서도 문제 해결 전략별로 학생의 답안을 '추출'하고 채점하는 방식은 교사가 최종 채점을 하기 전 초안으로 활용될 수 있다. 이는 채점자의 피로도를 줄이고 보다 일관된 채점을 지원하는 방안으로 고려될 수 있다.

Endnote

[1)]지도 학습은 기계 학습의 한 분야로 입력 데이터와 그에 대응하는 정답이 모델에 제공되며, 모델은 이 데이터를 활용하여 데이터의 패턴을 학습하는 것을 말한다.

[2)]자연어 처리란 인간의 음성이나 텍스트를 컴퓨터가 인식하고 처리하는 것을 말한다.

[3)]비지도 학습은 기계 학습의 한 분야로 입력 데이터만 모델에 제공되며 데이터 내의 구조나 패턴을 모델 스스로 찾아내어 학습하는 것을 말한다.

[4)]임베딩이란 단어, 문장 등을 기계가 이해할 수 있도록 숫자 형태인 벡터로 바꾸는 과정을 말한다.

[5)]MMLU 벤치마크는 언어 모델의 능력을 평가 하기 위해 다양한 학문적 주제와 실생활 시나리오에 기반한 문제를 제공하는 평가도구이다.

[6)]다운스트림 태스크는 사전 훈련된 AI 모델을 구체적인 실제 문제에 적용하여 평가하고 최적화하는 작업이다.

ORCID

Byoungchul Shin: https://orcid.org/0009-0000-1518-6846
Junsu Lee: https://orcid.org/0009-0009-5965-247X
Yunjoo Yoo: https://orcid.org/0000-0003-4769-0400

Conflict of Interest

The authors declare that they have no competing interests.

References

Baek, J. H., Shim, H. P., & Lee, D. W. (2023). *Analyzing and exploring the ability of chatbots for scoring student works in school science evaluation* (ORM 2023-30-7). KICE.

Batanero, C., Navarro-Pelayo, V., & Godino, J. D. (1997). Effect of the implicit combinatorial model on combinatorial reasoning in secondary school pupils. *Educational Studies in Mathematics, 32*(2), 181-199. https://doi.org/10.1023/A:1002954428327

Chang, S. C., & Kim, S. M. (2014). The defects of questions of descriptive assessment in elementary school mathematics and the suggestions for its improvement -focusing on the questions produced by Gyeonggi Provincial Office of Education. *Journal of Elementary Mathematics Education in Korea, 18*(2), 297-318.

Cho, M. J., Kim, M. R., Yoon, Y. G., & Shin, B. C. (2024). Development of an AI-integrated english writing class model for artificial intelligence literacy: Focused on prompt engineering. *Journal of Learner-Centered Curriculum and Instruction, 24*(6), 135-157. https://doi.org/10.22251/jlcci.2024.24.6.135

Choi, I. Y., & Cho, H. H. (2016). Eye movements in understanding combinatorial problems. *Journal of Educational Research in Mathematics, 26*(4), 635-662

Choi, S. K., & Park, J. I. (2023). A study on the development of automated Korean essay scoring model using random forest algorithm. *Brain, Digital, & Learning, 13*(2), 131-146. https://doi.org/10.31216/BDL.20230008

Chung, G. K., & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Chung, S. K., Lee, K. H., Yoo, Y. J., Shin, B. M., Park, M. M., & Han, S. Y. (2012). A survey of teachers' perspectives on process-focused assessment in school mathematics. *Journal of Educational Research in Mathematics, 22*(3), 401-427.

Ekin, S. (2023, March 4). *Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices.* https://doi.org/10.36227/techrxiv.22683919.v2

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*, 2629-2633. https://doi.org/10.1007/s10439-023-03272-4

Go, B. K., Shin, B. C. & Lim, C. I. (2024). Development of a math-AI convergence instructional model using a generative AI chatbot. *Journal of Educational Technology , 40*(1), 1-40. http://dx.doi.org/10.17232/KSET.40.1.1

Han, K. M., & Koh, S. S. (2014) An analysis of the mathematical errors on the items of the descriptive assessment in the equation of a circle. *The Mathematical Education, 53*(4), 509-524. https://doi.org/10.7468/mathedu.2014.53.4.509

Hwang, H. J., Na, G. S., Choi, S. H., Park, K. M., Lim, J. H., & Seo, D. Y. (2012). *Introduction to mathematics education.* Moonumsa.

Jin, K. A., Lee, B. C., Joo, H. M., & Shin, D. K. (2007). *Development of the KICE automated scoring program (II) (RRE 2007-4)*. KICE.

Jin, K. A., Lee, B. C., Shin, D. K., & Park, T. J. (2008). *The KICE automated scoring program (III) (RRE-2008-6)*. KICE.

Jin, K. A., Nam, M. H., Kim, M. H., Oh, S. C., Kim, M. J., & Joo, H. M. (2006). *A study on the development and introduction of an automated scoring program (RRI 2006-6)*. KICE.

Jung, H. D., Kang, S. P., & Kim, S. J. (2010). Analysis on error types of descriptive evaluations in the learning of elementary mathematics. *Journal of Elementary Mathematics Education in Korea, 14*(3), 885-905.

Jung, J. Y., Jo, H. M., Hwang, J. W., Moon, M. H., & Kim, I. J. (2023). *ChatGPT education revolution.* Porche.

Kapur, J. N. (1970). Combinatorial analysis and school mathematics. *Educational Studies in Mathematics, 3*, 111-127. https://doi.org/10.1007/BF00381598

Kim, M. J., Kim, Y. G., & J, I. C. (2009a). The study of factors of anxiety of permutation and coombination in high school. *Journal of the Korean School Mathematics, 12*(2), 261-279.

Kim, M. J., Kim, Y. G., & J, I. C. (2009b). The study of instruction on permutation and combination through the discovery method. *The Mathematical Education, 48*(2), 113-139.

Kim, M. K., Cho, M. K., & Joo, Y. R. (2012). A survey of perception and status about descriptive assessment - Focused on elementary school teachers in Seoul area. *Journal of Elementary Mathematics Education in Korea, 16*(1) 63-95.

Kim, R. Y., & Lee, M. H. (2013). Middle school mathematics teachers' perceptions of constructed-response assessments. *Journal of Educational Research in Mathematics, 23*(4), 533-551.

Kim, R. Y., Lee, M. H., Kim, M. K., & Noh, S. S. (2014). A comparison of elementary and middle school mathematics teachers' beliefs and practices in constructed-response assessment. *The Mathematical Education, 53*(1), 131-146.

https://doi.org/10.7468/mathedu.2014.53.1.131
Kim, S. R., Park, H. S., & Kim, W. S. (2007). Epistemological obstacles on learning the product rule and the sum rule of combinatorics. *The Mathematical Education, 46*(2), 193–205.
Kim, W. K., Hong, G. R., & Lee, J. H. (2011). Teaching and learning effects of structural-mapping used instruction in permutation and combination. *Communications of Mathematical Education, 25*(3), 607–627.
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (eds.), *Proceedings of the Advances in neural information processing systems 35 (NeurIPS 2022)* (pp. 22199–22213). https://doi.org/10.48550/arXiv.2205.11916
Kwon, O. N., Oh, S. J., Yoon J. E., Lee, K. W., Shin, B. C., & Jung W. (2023). Analyzing mathematical performances of ChatGPT: Focusing on the dolution of national sssessment of educational achievement and the college scholastic ability test. *Communications of Mathematical Education, 37*(2), 233–256. https://doi.org/10.7468/jksmee.2023.37.2.233
Lee, J. S. (2024). *A exploration of unsupervised learning-based grading aid methods for mathematical descriptive assessment* [Master's thesis, Seoul National University]. http://www.dcollection.net/handler/snu/000000182358
Lee, K. C. (2021). *Do it! Natural language processing with Bert and GPT*. EasysPublishing.
Lee, K. S., Rim, H. M., Choi, I. S., & Kim, S. K. (2021). *The theory and practice of mathematics education assessment.* Kyowoo.
Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638
Ministry of Education. (2022). *Mathematics curriculum*. Ministry of Education Notice 2022-33 [supplement 8]. Ministry of Education.
Na, G. S., Park, M. M., Park, Y. J., & Lee, H. C. (2018). A study on mathematical descriptive evaluation – Focusing on examining the recognition of mathematics teachers and searching for supporting way. *School Mathematics, 20*(4), 635–659. https://doi.org/10.29275/sm.2018.12.20.4.635
National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. NCTM.
Noh, E. H., Kim, M. H., Sung, K. H., Kim, H. S., & Jin. K. Y. (2013). *Refinement and pilot implementation of an automated scoring program for open-ended items in large-scale assessments* (Research Report RRE 2013-5). Korea Institute for Curriculum and Evaluation.
Noh, E. H., Lee, S. H., Lim, E. Y., Sung, K. H., & Park, S. Y. (2014). *Development and practical validation of an automated scoring program for Korean open-ended items* (Research Report RRE 2014-6). Korea Institute for Curriculum and Evaluation.
Noh, E. H., Shim, J. H., Kim, M. H., & Kim, J. H. (2012). *A study on automatic Scoring methods for large-scale assessments of open-ended items* (Research Report RRE 2012-6). Korea Institute for Curriculum and Evaluation.
Noh, E. H., Song, M. Y., Park, J. I., Kim, Y. H., & Lee, D. K. (2016). *Advanced refinements and application of automated scoring system for Korean large-scale assessment* (Research Report RRE 2016-11). Korea Institute for Curriculum and Evaluation.
Noh, E. H., Song, M. Y., Sung. K. H., & Park, S. Y. (2015). *Development and implementation of an automated scoring program for sentence-level open-ended items in Korean* (Research Report RRE 2015-9). Korea Institute for Curriculum and Evaluation.
Noh, S. S., Kim, M, K., Cho, S. M., Jeong, Y. S., & Jeong. Y. A., (2008). A study of teachers' Perception and status about descriptive evaluation in secondary school mathematics. *Journal of the Korean School Mathematics, 11*(3), 377–397.
Oh, S. J. (2023). Effective ChatGPT prompts in mathematical problem solving: Focusing on quadratic equations and quadratic Functions. *Communications of Mathematical Education, 37*(3), 945–967. https://doi.org/10.7468/jksmee.2023.37.3.545
OpenAI (2023). *GPT-4 technical report.* https://arxiv.org/abs/2303.08774
OpenAI (2024, Feb, 1). *Six strategies for getting better results. OpenAI API*. https://platform.openai.com/docs/guides/prompt-engineering
Pang, J. S., Kim, S. H., An, H. J., Chung, J. S., & Kwak, G. W. (2023). Challenges faced by elementary teachers in implementing the five practices for effective mathematical discussions. *The Mathematical Education, 62*(1), 95–115. https://doi.org/10.7468/mathedu.2023.62.1.95
Park, J. I., Lee, S. H., Song, M. H., Lee, M. B., Lee, M. J., & Choi, S. K. (2022). *A study on the development of automated scoring method for computer-based essay and short answer question type assessment I* (Research Report RRE 2022-6). Korea Institute for Curriculum and Evaluation.
Park, J. I., Lee, S. H., Song, M. H., Lee, M. B., Lee, M. J., & Choi, S. K. (2023). *A study on the development of automated scoring method for computer-based essay and short answer question type assessment II* (Research Report RRE 2023-7). Korea Institute for Curriculum and Evaluation.
Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). Chatbots put to the test in math and logic items: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI, 4*(4), 949–969. https://doi.org/10.3390/ai4040048

Seo, J. Y., Nam, M. H., Kim, S. Y., Lee, W. S., Choi, M. S., Hong, S. J., & Kwon, Y. M. (2010). *Study on the activation of performance assessment for the development of creativity and cultivation of character* (Research Report CRE 2010–16). Korea Institute for Curriculum and Evaluation.

Seong, J. W., & Shin, B. C. (2023). Exploring the feasibility of automatic scoring of written test using ChatGPT: Focusing on the world geography written test. *Journal of the Association of Korean Geographers, 12*(3), 415–432. https:// doi.org/10.25202/JAKG.12.3.3

Seong, T. J., & Kwon, O. N. (1999). Future directions and perspectives for performance assessment in mathematics. *Journal of the Korea society of Educational Studies in Mathematics School Mathematics, 1*(1), 217–234.

Song, M. Y., Noh, E. H., & Sung, K. H. (2016). Analysis on the accuracy of automated scoring for Korean large-scale assessments. *The Journal of Curriculum and Evaluation, 19*(1), 255–274. https://doi.org/10.29221/jce.2016.19.1.255

Sriraman, B., & English, L. D. (2004). Combinatorial mathematics: Research into practice. *Mathematics Teacher, 98*(3), 182–191. https://doi.org/10.5951/MT.98.3.0182

Sung, J. H. (2023). Analysis of functions and applications of intelligent tutoring system for personalized adaptive learning in mathematics. *The Mathematical Education, 62*(3), 303–326. https://doi.org/10.7468/mathedu.2023.62.3.303

Taulli, T. (2023). *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business (e-book)*. Berkeley. https:// doi.org/10.1007/978-1-4842-9367-6_1

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv*:2203.11171. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *In Proceedings of the Advances in Neural Information Processing Systems 35* (pp.24824–24837). https://doi.org/10.48550/arXiv.2201.11903

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate item solving with large language models. *In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. https://doi.org/10.48550/arXiv.2305.1060

Yoon, S. Y., Eva Miszoglad, & Lisa R. Pierce. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. *arXiv preprint arXiv*:2310.06505. https://doi.org/10.48550/ arXiv.2310.06505

MathEdu Journal of the Korean Society of Mathematical Education Series A

Research Article

The Mathematical Education 2024;63(2):187-207
https://doi.org/10.7468/mathedu.2024.63.2.187
pISSN 1225-1380, eISSN 2287-9633

프롬프트 엔지니어링을 통한 GPT-4 모델의 수학 서술형 평가 자동 채점 탐색: 순열과 조합을 중심으로

신병철[1], 이준수[2]*, 유연주[3]

[1]수원외국어고등학교 교사
[2]화홍고등학교 교사
[3]서울대학교 교수

*교신저자: 이준수(jnsulee@gmail.com)

초 록

본 연구에서는 GPT-4 기반의 ChatGPT를 활용한 서술형 평가 문항의 자동 채점 가능성을 탐색하기 위해 교사와 GPT-4 기반의 ChatGPT의 채점 결과를 비교, 분석하였다. 이를 위해 학생평가지원포털에 있는 고등학교 1학년 순열과 조합 단원에서 3개의 서술형 문항을 선정하였다. 문항 1, 2는 문제 해결 전략이 1가지인 문항이고, 문항 3은 문제 해결 전략이 2가지 이상인 문항이었다. 8년 이상의 교육 경력이 있는 교사 2명이 학생 204명의 답안을 채점하고, GPT-4 기반의 ChatGPT의 채점 결과와 비교하였다. 문항별로 Few-Shot-CoT, SC, 구조화, 반복 프롬프트 기법 등을 활용하여 채점을 위한 프롬프트를 구성하였고, 이를 GPT-4 기반의 ChatGPT에 입력하여 채점하였다. 채점 결과, 문항 1, 2는 교사의 채점 결과와 GPT-4의 채점 결과 사이에 강한 상관관계를 충족하였다. 문제 해결 전략이 2가지인 문항 3은 먼저 채점 전 학생 답안을 문제 해결 전략별로 분류하는 프롬프트를 GPT-4 기반의 ChatGPT에 입력하여 답안을 분류하였다. 이후 유형별로 채점 프롬프트를 적용하여 GPT-4 기반의 ChatGPT에 입력하여 채점하였고, 채점 결과 역시 교사의 채점 결과와 강한 상관관계가 나타났다. 이를 통해 프롬프트 엔지니어링을 활용한 GPT-4 모델이 교사의 채점을 보조할 수 있는 가능성을 확인하였으며 본 연구의 한계점 및 향후 연구 방향을 제시하였다.

주요어 프롬프트 엔지니어링, 자동 채점, 서술형 평가, GPT-4기반의 ChatGPT, 생성형 인공지능, 순열과 조합